Increasing Object-Level Reconstruction Quality in Single-Image 3D Scene Reconstruction

Anna Ribic Antonio Oroz Meikel Kokowski Franz Srambical
Technical University of Munich
{firstname}.{lastname}@tum.de

Abstract

Panoptic 3D scene reconstruction describes the joint task of geometric reconstruction, 3D semantic segmentation, and 3D instance segmentation. A multitude of tasks in robotics, augmented reality and human-computer interaction rely on this comprehensive understanding of 3D scenes. Extending prior work which performs panoptic 3D scene reconstruction from a single RGB image, our proposal aims to enhance the visual clarity and discernibility of the generated geometry using an additional object-level generative approach. Leveraging an existing 3D asset generation framework, we integrate object-level reconstruction into the 3D scene, yielding superior aesthetic quality.

1. Introduction

While humans can easily infer the 3D structure as well as the complete (panoptic) semantics of a scene from a single image, this task has been a longstanding challenge in the field of computer vision. The task fundamentally prerequisites learning a strong prior of the 3D world. Traditional methods have made significant strides, from generating geometrically coherent structures [11, 40] to learning different instance semantics [17, 27, 36]. More recent approaches directly learn the 3D panoptic semantics as a whole [8, 50], yet they fall short in capturing the intricate details and nuances at the object level. This paper introduces a novel approach to bridge this gap by integrating a specialized object-level model into the reconstruction process, thereby leveraging the specialized model's object-priors.

Our approach models panoptic 3D reconstruction as a two-stage problem. We first use the model of Dahnert et al. [8] to create an initial reconstruction. Then, we leverage the instance masks to extract the object geometries out of the reconstructed scene. We input each of the extracted objects along with cropped images from the scene and text labels into a diffusion model [5] to refine the rough object-level geometries. Finally, we integrate the refined object geometries back into the initial scene reconstruction to obtain a complete and refined panoptic 3D scene reconstruction.

In summary, our main contributions are as follows:

- We propose a novel approach to panoptic 3D reconstruction involving an inference pipeline that leverages object-level reconstruction models to refine the output of a 3D scene reconstruction backbone.
- We generate a new synthetic 3D-Front [15] dataset comprising over 24,000 samples, each annotated with both 2D and 3D ground truth data.
- We qualitatively demonstrate the effectiveness of our approach on the 3D-Front [15] dataset, showing significant improvements over the state-of-the-art.
- We show that fine-tuning SDFusion [5] on the input scene's object distribution (in our case the 3D-Future dataset [16]) significantly improves the quality of the refined objects.
- We introduce a conceptually simple yet effective method for shape alignment, which outperforms rigid alignment methods in our experiments.

2. Related Work

2D panoptic segmentation 2D panoptic segmentation merges semantic and instance segmentation, providing detailed pixel-level parsing of images, capturing both general categories (semantic segmentation) and individual object identities (instance segmentation) [25]. Since the original task formulation by Kirillov et al. [25], a number of works have been proposed to solve the task [2–4, 26, 28–30, 35, 43, 44, 47–49], while more recent approaches [23] try to unify image segmentation in its entirety.

Single-view 3D reconstruction The work by Snavely et al. [41] was the first notable attempt at reconstructing 3D scenes from unordered photo collections. Since then, the field of image-based 3D reconstruction has seen a number of advancements, culminating in the task of single-view 3D reconstruction [6, 11, 22, 33, 36, 40, 45].

Shape priors Wu et al. [46] note that the task of single-view 3D reconstruction is non-deterministic, as there are many 3D shapes that can explain a given single-view input, and propose to use shape priors to shape the solution space such that the reconstructed shapes are realistic, but not necessarily the ground truth. Our approach extends the idea of Wu et al. [46] to entire 3D scenes.

3D scene understanding The task of 3D scene understanding and panoptic reconstruction is analogous to its 2D counterpart and aims to infer the 3D structure and semantics of a scene, including the 3D layout, object instances, and their 3D shapes from images [8] or noisy geometry [20, 21]. Dahnert et al. [8] propose a method – henceforth called *Panoptic 3D* – to jointly solve the tasks of 3D scene understanding and single-view 3D reconstruction by lifting features produced by a 2D backbone into a 3D volume of the camera frustrum, and jointly optimizing for geometric reconstruction as well as semantic and instance segmentation.

Modality-conditioned shape generation 3D generative models represent objects in a variety of modalities, including point clouds [1, 32], occupancy grids [33], meshes [34], and signed distance functions [38]. Furthermore, these models can also be distinguished by the type of input they take, such as incomplete shapes [10], images [14], text [31, 51], or other modalities [52]. Notably, Cheng et al. [5] propose *SDFusion*, a 3D object reconstruction method conditioned on images, text and geometrical input.

Datasets Notable datasets in the field of panoptic 3D reconstruction include ScanNet [9] and Replica [42], which provide rich annotations for scene understanding tasks. Another such dataset, 3D-Front [15], provides comprehensive coverage of indoor scenes while offering detailed geometric reconstructions as well as semantic and instance segmentation annotations. The synthetic 3D dataset contains 6,801 mid-size apartments with 18,797 rooms populated by 3D shapes from the 3D-Future [16] dataset. The dataset's high-quality data acquisition process ensures accurate representations, establishing it as a valuable resource for advancing research in 3D panoptic reconstruction.

In an effort to refine the panoptic reconstruction model, we compiled a custom dataset comprising over 24,000 samples. Leveraging the diverse scenes of the 3D Front dataset, we use BlenderProc [12] for randomly sampling camera poses and 2D rendering. Utilizing the C++ pipeline from Dahnert et al. [8], we generate annotated 3D geometry within the respective camera frustum.

3. Method

3.1. Initial Panoptic Scene Reconstruction

We leverage Panoptic 3D [8] to predict the camera frustum geometry $\mathbf{X}_{P_{\text{geom}}}$ as well as associated 3D semantic and instance labels $\mathbf{X}_{P_{\text{sem}}}$, $\mathbf{X}_{P_{\text{instance}}}$ within the image. Said model yields both 2D and 3D representations of detected objects and does so by employing a ResNet-18 [18] encoder for feature extraction from the input image. Subsequently, both a depth encoder and a Mask R-CNN [19] are applied to the ResNet-18 encoder features to predict both a 2D depth map and a 2D instance mask. During training, we learn the 2D output utilizing proxy losses for both depth estimation (L_d) and instance segmentation (L_i).

The depth map facilitates the backprojection of features into a sparse volumetric grid, while the 2D instance mask is propagated to serve as a seed for the 3D instance mask prediction. Finally, a 3D U-Net [7] processes the sparse backprojection to forecast occupancy, distance field, and both semantic and instance labels for each individual occupancy within the grid.

In addition to the proxy losses, binary cross-entropy is used on the occupancy prediction at different hierarchy levels and an l_1 loss is employed on the distance field at the final hierarchy level. The total loss can be formalized as

$$\mathcal{L} = w_d\mathcal{L}_d + w_i\mathcal{L}_i + \sum_h (w_g\mathcal{L}_g^h + w_s\mathcal{L}_s^h + w_o\mathcal{L}_o^h), \quad (1)$$

where $\mathcal{L}_g^h, \mathcal{L}_s^h, \mathcal{L}_o^h$ represent the geometry as well as 3D semantic and instance label losses at different hierarchy levels, and $w_{x\in\{d,i,g,s,o\}}$ being weighting factors.

At inference time, we use the 2D instance mask to extract RGB crops $\mathbf{I}_{\text{crop}}$ of the input image, and the 3D instance mask to extract the corresponding 3D geometry $\mathbf{X}_{P_{\text{geom, crop}}}$. The extracted image, geometry and the semantic label are subsequently input into the object-level reconstruction model for shape reconstruction.

3.2. Object-Level Reconstruction

We use SDFusion [5] for object shape reconstruction, which expects a signed distance field as its primary input, and additionally leverages an RGB image and a textual representation as conditional inputs to guide the reconstruction process. To this end, SDFusion employs task-specific encoders ([13, 39]) to get image and text embeddings, while simultaneously embedding the 3D shape into a latent space using a pre-trained vector quantized variational autoencoder (VQ-VAE) [37]. At training time, noise is introduced to the shape latent via forward diffusion, which is followed by a concatenation of the conditional embeddings. This serves as input to the 3D U-Net [7] denoising network which reconstructs the latent code. Within the denoising U-Net, cross-attention is applied along the concatenated latent code to modulate the denoising process. Ultimately, the VQ-VAE decoder reconstructs the shape.

At inference time, we use SDFusion to output a refined object geometry $\mathbf{X}_S$ for every object-level geometry extraction $\mathbf{X}_{P_{\text{geom, crop}}}$, leveraging the image crop $\mathbf{I}_{\text{crop}}$ and the corresponding semantic label $\mathbf{X}_{P_{\text{sem}}}$.

3.3. Object-Level Shape Alignment

The inference output of SDFusion is front-facing and might not align with the object's orientation in the original 3D scene. Thus, to adequately replace the original objects with the refined ones, we employ a custom registration algorithm to ensure proper alignment of the reconstructed objects within the scene. This process consists of 3 key steps:

1. **Floor alignment:** To establish a common frame of reference and facilitate subsequent re-orientation, we align the reconstructed object with the floor plane of the 3D scene.
2. **Rotational optimization:** Following floor alignment, the object is systematically rotated to 16 discrete, uniformly distributed positions around its y-axis, convering a diverse set of potential orientations.
3. **Selection based on similarity:** The final step involves selecting the orientation that minimizes the per-point difference between the reconstructed object (mesh) and the corresponding elements within the scene (point) utilizing trimesh. This metric serves as a quantitative measure of alignment accuracy.

4. Results

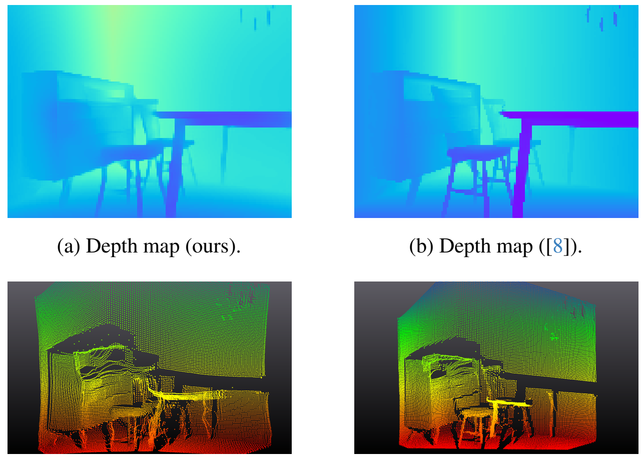
Our experiments demonstrate the efficacy of our method in enhancing the visual representation of objects within reconstructed 3D scenes. Figure 1 (TODO FIGURE 1 IS WRONG!) shows that the output of Panoptic 3D exhibits visual artificats and irregularities, while our approach yields smooth surfaces, improving visual aesthetics. Notably, we observe sensitivity of our alignment procedure to the quality of the instance masks generated by Panoptic 3D. Although the refined objects maintain smoothness even in bad quality mask predictions, our alignment algorithm occasionally fails in unfavourable conditions, resulting in object intersections, as illustrated in Fig. 3.

Panoptic Reconstruction Training We leverage our synthesized dataset to refine the training of the panoptic reconstruction model proposed by Dahnert et al. [8]. Initially, we pre-train the 2D encoder, depth estimation and 2D instance prediction using the ADAM optimizer [24] using a batch size of 1 and learning rate 1e-4 for 570k iterations.

The evaluation results for our 2D model compared to the pre-trained model from Dahnert et al. [8] are presented in Tab. 1. As illustrated in Fig. 1, our approach shows performance comparable to the pre-trained model. However, it encounters challenges in generating completely clear depth results, occasionally displaying some irregularities. Despite our efforts, limitations such as time constraints and the relatively small size of our dataset hindered our ability to train a 3D model that achieves performance on par with the pre-trained counterpart. We refer to the future work section in this regard.

	Depth	Box Class.	Box Regress.
Dahnert et al. [8]	0.23	3.39	**0.092**
Ours	**0.196**	**1.3**	0.149

Table 1. Results for joint training of the 2D encoder, depth estimation and 2D instance prediction. For depth we report the ℓ_1 distance between the predicted and ground-truth depth maps. Additionally we report the ℓ_1 distance for the regressed 2D boxes and a CE-loss on the box classification.



(a) Depth map (ours). (b) Depth map ([8]).

(c) Geometry from depth (ours). (d) Geometry from depth ([8]).

Figure 1. 2D results from the Panoptic 3D model. Our re-training results (left) vs. results from Dahnert et al. [8] (right).

SDFusion Fine-tuning In order to align SDFusion to the shape distribution of 3D-Front, we fine-tune the model on a subset of the 3D-Future dataset [16], which contains the objects used in the scenes of 3D-Front. We train the model for 12k steps using the original hyperparameters from Cheng et al. [5] but with a batch size of 32 (see Figure 2).

Inference Pipeline TODO add visual results

5. Conclusion, Limitations & Future Work

Conclusion In summary, our method has demonstrated the effectiveness of combining reconstruction techniques with a object-level generation framework to significantly enhance the aesthetic quality of 3D instances within a reconstructed

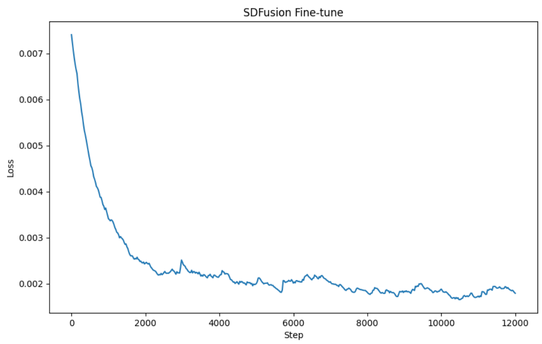
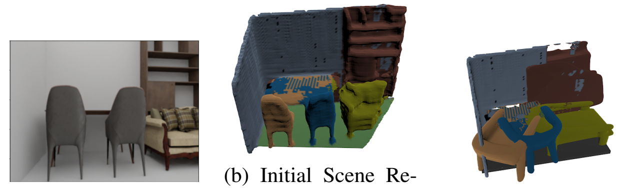


Figure 2. Loss curve of our SDFusion fine-tune (smoothed).

scene. Moreover, our approach facilitates a straightforward process for generating virtual environments from single images, with potential applications in gaming, virtual reality, and augmented reality settings. An additional advantage of utilizing a multi-modal diffusion model instead of directly leveraging high-quality 3D objects, such as CAD models, is the ability to customize inputs, enabling fine-grained control over the reconstructed scene.

Limitations While our results show promise, it's important to acknowledge certain limitations of our method. Firstly, concerns arise regarding object detection. While the panoptic reconstruction model doesn't necessarily need to detect an instance to reconstruct its approximate shape, SDFusion is only applied to detected objects. As a result, undetected objects and instances erroneously identified as part of another object may be entirely omitted from the refinement process. Moreover, noisy instance segmentations pose another challenge, as our scale and position estimations are contingent upon the predicted instance labels. Hence, instance segmentations that include parts of other instances or elements of the surrounding environment can lead to the creation of larger, misplaced refined instances. A concrete example of this phenomenon is presented in Fig. 3.

Future Work To overcome these limitations, we suggest two avenues for future research. The first direction involves implementing end-to-end training with adapted loss functions, aimed at penalizing misidentified instances more effectively. This approach could enhance the accuracy of instance segmentation and reduce the incidence of missing or misattributed objects in the reconstructed scene. Secondly, refining the merging process by integrating a pose estimation network can be utilized to enhance object alignment and scaling. Another promising avenue for exploration involves guiding object-level reconstruction with more detailed descriptive inputs. By incorporating these object descriptions during inference, inspired by the findings in the SDFusion paper, we can potentially generate more tailored and contextually relevant reconstructions.



(a) Input Image (b) Initial Scene Reconstruction (c) Final Output

Figure 3. Our method struggles with missing and ambiguous semantic/instance labels. As the initial instance segmentation fails to identify the 'Table' object, the corresponding geometry is entirely absent from our reconstruction. Furthermore, since the 'Chair' instance masks overlap with the table geometry, our method generates corresponding instances with incorrect scale and pose.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 2

[2] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J. Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 909–919, 2023. 1

[3] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020.

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 1

[5] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 1, 2, 3

[6] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 1

[7] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2

[8] Manuel Dahnert, Ji Hou, Matthias Nießner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. *Advances in Neural Information Processing Systems*, 34:8282–8293, 2021. 1, 2, 3

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[10] Angela Dai, Charles Ruizhongtai Qi, and Matthias Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5868–5877, 2017. 2

[11] Maximilian Denninger and Rudolph Triebel. 3d scene reconstruction from a single viewport. In *European Conference on Computer Vision*, pages 51–67. Springer, 2020. 1

[12] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Knauer, Klaus H. Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[14] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 2

[15] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 1, 2

[16] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 1, 2, 3

[17] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9785–9795, 2019. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[20] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019. 2

[21] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 2

[22] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Perspectivenet: 3d object detection from a single rgb image via perspective points. *Advances in neural information processing systems*, 32, 2019. 1

[23] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2989–2998, 2023. 1

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[25] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1

[26] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J. Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12871–12881, 2022. 1

[27] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 260–277. Springer, 2020. 1

[28] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 1

[29] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 214–223, 2021.

[30] Zhiqi Li, Wenhai Wang, Enze Xie, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo, and Tong Lu. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 1

[31] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2022. 2

[32] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 2

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 1, 2

[34] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019. 2

[35] Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021. 1

[36] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1

[37] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 2

[38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[40] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2172–2182, 2019. 1

[41] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 1

[42] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[43] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Standalone axial-attention for panoptic segmentation. In *European conference on computer vision*, pages 108–126. Springer, 2020. 1

[44] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5463–5474, 2021. 1

[45] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 1

[46] Jiajun Wu, Chengkai Zhang, Xiuming Zhang, Zhoutong Zhang, William T Freeman, and Joshua B Tenenbaum. Learning shape priors for single-view 3d completion and reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 646–662, 2018. 2

[47] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 1

[48] Qihang Yu, Huiyu Wang, Dahun Kim, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2560–2570, 2022.

[49] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *arXiv preprint arXiv:2308.02487*, 2023. 1

[50] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3d: A universal model for panoptic 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9256–9266, 2023. 1

[51] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *arXiv preprint arXiv:2306.17115*, 2023. 2

[52] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4176–4186, 2021. 2